SUMMARY OF CLAIMED SUBJECT MATTER

The present invention contains claims to a method for discovering patterns in a set of sequences of symbols (claims 35, 42 and 44), to a computer-readable medium containing a data structure useful by a computer system in the practice of the method (claim 66), and to a computer-readable medium containing instructions for controlling a computer system to perform the method (claim 68).

In all of its various aspects the present invention is directed to the identification of existing patterns in a set of "k" number of sequences. The k number of sequences is termed a "k-tuple". The k number of sequences form part of an overall set of "w" number of sequences. Each of the w sequences has a given length, but the sequences need not be the same length. A "pattern" is a distributed substring of elements that occurs in at least two sequences in a set of sequences (Page 7, lines 36-38).

The basic steps that comprise the core of the method of the present invention may be understood from the following discussion of a "two-tuple" ($k = 2$) of sequences S_1 and S_2.

Each member element of a sequence is represented by an alphabetic symbol. Page 7, lines 30-31 and 33-34 show the two representative sequences S_1 and S_2. Each symbol occupies a given location in a sequence. This location is termed the symbol's "position index". The pairing of a symbol and its position index identifies a unique symbol at a unique location in a sequence.

The first step of the method is to create for each sequence a table of ordered (symbol, position index) pairs. For instance, in the sequence S_1 the symbol "L" occurs at position indices 18 and 46 while the symbol "K" occurs at position indices 20, 25, 34 and 35. In the sequence S_2 the symbol "L" occurs at position indices 6, 23 and 30 and the symbol "K" occurs at position indices 8, 10, 14 and 32. The creation of this table of ordered pairs corresponds to step (a) of claim 68.

The association of each symbol and its position index is used to form a "master offset table" for each sequence. Figure 1 shows two master offset tables for the "two-tuple" of sequences S_1 and S_2. Each master offset table groups, for each symbol, the position in the sequence occupied by each occurrence of that symbol. The master offset tables are the first data structures recited in claim 66 and are created by step (b) of claim 68.

Thus, in the master offset table of Figure 1 for the sequence S_1 the position indices "18" and "46" are listed under the symbol "L" while position indices "20", "25", "34" and "35" are listed under the symbol "K". Similarly, for the sequence S_2 position indices "6", "23" and

"30" are listed for the symbol "L" and position indices "8", "10", "14" and "32" are listed for the symbol "K".

Next, the difference-in-position between each occurrence of a symbol in one of the sequences and each occurrence of that same symbol in the other sequence is determined. This determination is facilitated by concatenating the two sequences. This is described at Page 10, line 15. A table, termed a "pattern map" (page 9, lines 32-33) or a "tuple-table" (page 30, line 25 through page 31, line 15), is formed in which each row in the table represents a single value of "difference in position" (page 9, line 20 through page 10, line 6). This pattern map is the "k-tuple table data structure" recited in claim 66 and is produced by step (c) of claim 68.

Figures 2A and 2B depict the pattern map for the two-tuple of sequences S_1 and S_2. Since sequence S_1 contains 47 characters and the sequence S_2 contains 54 characters the pattern map is 101 rows in depth (rows numbered "0" through "100"). For each given value of a difference-in-position (the value being termed the "row index") the table lists the position of each symbol in the first sequence that appears again at a spacing corresponding to that difference-in-position value. The column of row indices in the table of Figures 2A and 2B is referred to as "the primary column" in claim 66.

Consider the symbol "R" listed in the master offset table for the sequence S_1 (at position index "44") and the position indices for the same symbol "R" as listed in the master offset table for the sequence S_2 (position indices "7", "21", "31"). From the master offset tables and the concatenation of the sequences S_1 and S_2 at page 7 it may be determined that:

- from the occurrence of the symbol "R" in the first sequence,

--the first occurrence of the symbol "R" in the second sequence is spaced ten places;

–the second occurrence of the symbol "R" in the second sequence is spaced twenty-four places; and

--the third occurrence of the symbol "R" in the second sequence is spaced thirty-four places.

The pattern map of Figures 2A and 2B thus lists the position index "44" (corresponding to the symbol "R") on row indices (difference-in-position values in the primary column) of "10", "24" and "34".

The symbols collected for any row index (each value of difference-in-position) define a parent pattern in the first sequence that is repeated in the second sequence.

Consider the discussion at page 11, line 24 through page 12, line 30 for the row index value "35" in the primary column in the pattern map (or "two-tuple-table") of Figure 2A. This row index value identifies the pattern corresponding to the symbols at position indices "18", "20", "21", "30", "39" and "40". [The value "6" in the symbol count column to the immediate right of the colon on Figure 2A (page 10, line 37 through page 11, line 2) indicates that there are six symbols in the pattern.] Figures 2A and 2B thus show a sorted k-tuple table that is created by row-sorting the k-tuple table by the position indices in the primary column. This sorted k-tuple table is the "sorted k-tuple table data structure" recited in claim 66 and is included in step (d) of claim 68.

By consulting sequence S_1 the position indices "18", "20", "21", "30", "39" and "40" respectively correspond to the symbols "L", "K", "V", "V", "P", "H".

The collected symbols corresponding to a difference-in-position value "35" thus identifies the pattern occurring in the first sequence S_1 as:

"L . KV . . . . . V . . . . . PH"

that also appears in the second sequence S_2 (page 12, line 16), where the dots indicate placeholders in the pattern (page 12, lines 19-21). This is recited in step (e) of claim 68.